Using Machine Learning in Adversarial Environments
173037

Year 2 of 3
Principal Investigator: W. L. Davis iv / 01461
Investment Area(s): Defense Systems and Assessments

Project Purpose:

Intrusion/anomaly detection systems are among the first lines of cyber defense. Commonly, they either use signatures or machine learning (ML) to identify threats, but fail to account for sophisticated attackers trying to circumvent them. We propose to embed machine learning within a game theoretic framework that performs adversarial modeling, develops methods for optimizing operational response based on ML, and integrates the resulting optimization codebase into the existing ML infrastructure developed by the Hybrid LDRD. Our approach addresses three key shortcomings of ML in adversarial settings: 1) resulting classifiers are typically deterministic and, therefore, easy to reverse engineer; 2) ML approaches only address the prediction problem, but do not prescribe how one should operationalize predictions, nor account for operational costs and constraints; and 3) ML approaches do not model attackers' response and can be circumvented by sophisticated adversaries. The principal novelty of our approach is to construct an optimization framework that blends ML, operational considerations, and a model predicting attackers reaction, with the goal of computing optimal moving target defense. One important challenge is to construct a realistic model of an adversary that is tractable, yet realistic. We aim to advance the science of attacker modeling by considering game-theoretic methods, and by engaging experimental subjects with red teaming experience in trying to actively circumvent an intrusion detection system, and learning a predictive model of such circumvention activities. In addition, we will generate metrics to test that a particular model of an adversary is consistent with available data.

Little research has been conducted to date on providing a mathematical framework for rigorously analyzing the impact of defensive postures (e.g., network configurations) on the interactions between cyber defenders and adversaries. This work is novel and risky, yet, if successful, could serve as the catalyst for extending theoretical understanding of interactions towards practical decisions that will improve enterprise-wide computer security.

Refereed Communications:
Laszka, A.; Vorobeychik, Y.; Koutsoukos, X.. "Optimal Personalized Filtering Against Spear-Phishing Attacks". AAAI Conference on Artificial Intelligence, North America, feb. 2015. Available at: <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9917>. Date accessed: 30 Sep. 2015.

Li, Bo; Vorobeychik, Yevgeniy. "Scalable Optimization of Randomized Operational Decisions in Adversarial Classification Settings". Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, San Diego, California, USA, May 9-12, 2015.

Li, Bo; Vorobeychik, Yevgeniy. "Feature Cross-Substitution in Adversarial Classification". Advances in Neural Information Processing Systems, 2014 pp. 2087--2095.